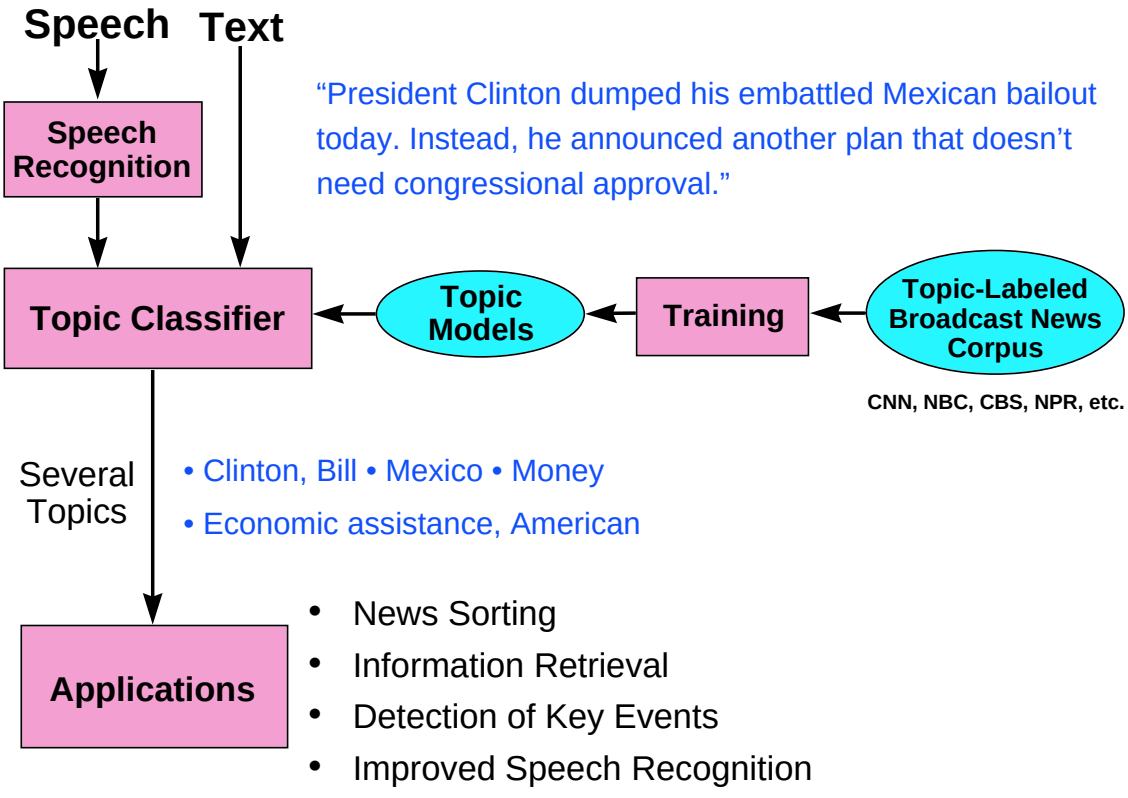
Topic Indexing of Broadcast News

Richard Schwartz

BBN Systems and Technologies

28 March 1997

Topic Indexing

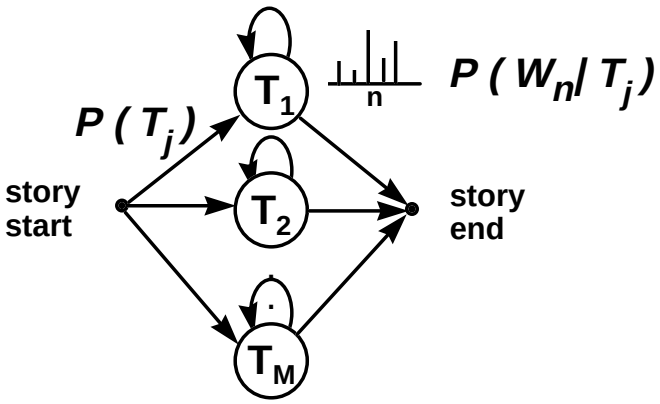


Primary Source Media Corpus

- Each news story is annotated with several detailed topics
  - from 1 to 13 topics
  - average 4.5 topics / story
- Example: story about Bill Clinton talking about loans to Mexico is labeled with
  - Bill Clinton
  - Mexico
  - money
  - American economic assistance
- The topic set is open
  - 5,000 topics in one year
  - 9,000 topics in four years

Traditional Generative Topic Model

- First, choose a topic, T_j
- For each important word:
  - choose the word according to $P(W_n | T_j)$
  - unimportant words can be inserted randomly



- Model assumes there is only one true topic per story.
- All words in a story are assumed related to the topic.

Recognition Score for a Topic

- Use Bayes' rule and assume words are independent

$$P(T_j \mid \text{Words}) = P(T_j) \frac{P(\text{Words} \mid T_j)}{P(\text{Words})} \qquad \text{Bayes' Rule}$$

$$\approx \Pr(T_j) \prod_t \frac{P(W_t \mid T_j)}{P(W_t)} \qquad \text{Independence}$$

- Estimation noise (e.g., unobserved words for a topic) is a problem.
  - Solution 1: Discard "unimportant" words (e.g., words with low mutual information)
  - Solution 2: Smooth estimates with unconditioned model

$$\Pr(W_t \mid T_j) = \alpha \Pr(W_t \mid T_j) + (1-\alpha) \Pr(W_t)$$

Problems with Traditional Model

- Model assumes there is only one true topic per story.
  - but typical stories have several topics

Story:

"**President Clinton dumped** his **embattled Mexican bailout** today. Instead, he **announced** another plan that doesn't need **congressional approval**."

Annotated Topics:

Clinton, Bill; Mexico; Money; Economic assistance, American

  - This causes the distributions of words to overlap among topics

- Model assumes all **important** words are related to all **topics.**
  - But, for example, the word **Clinton** does not imply the topic **Mexico** in all stories.
  - Words from one topic are taken as negative evidence for another topic

Problems with Traditional Models (Cont.)

- Because distributions of words are estimated by simple counting, common words have higher probability than real keywords.

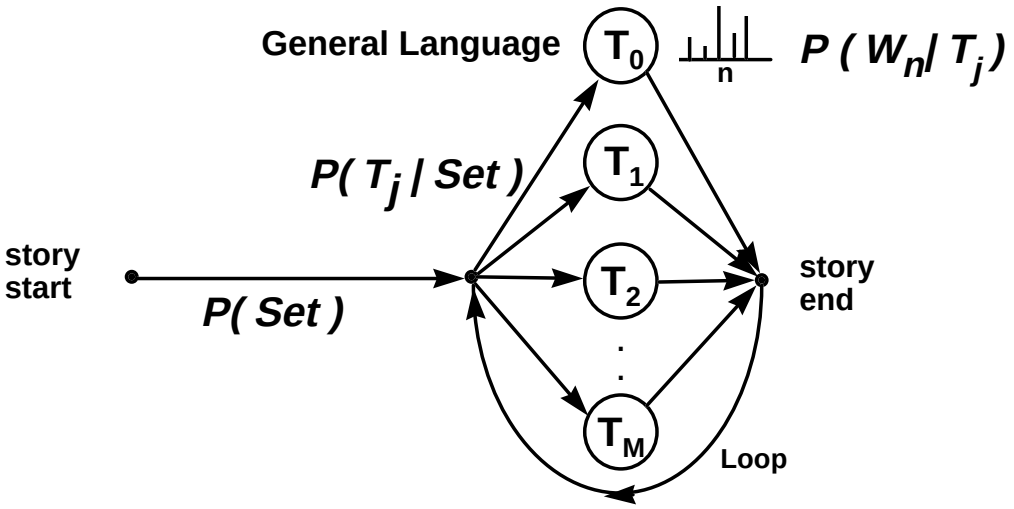
Rank	Word	P(W | T= "Clinton, Bill")
1	president	0.013
2	go	0.011
3	think	0.010
4	Clinton	0.009
5	say	0.008

New Model

- Assume each story has several topics.
- Some words (keywords) are related to each topic.
  - Different words may be related to different topics.
- All words are related to some topic
  - Most words are related to general language (T_0).

New Generative Model of Topic

- First, choose a Set of topics, $T_0 ... T_M$
- Normalize $P (T_j | Set): \sum_j P (T_j | Set) = 1$
- For each word:
  - Choose a topic according to $P (T_j | Set)$
  - Choose a word according to output distribution $P (W_n | T_j)$
  - Loop



Recognition Score for a Topic

$$P (Set | Words) = P (Set) \frac{P (Words | Set)}{P (Words)}$$ Bayes' Rule

$$\frac{P (Words | Set)}{P (Words)} \approx \prod_t \frac{P (W_t | Set)}{p (W_t)}$$ Independence

$$P (W_t | Set) = \sum_j P (T_j | Set) \; p (W_t | T_j)$$

$$P (Set | Words) \approx P (Set) \prod_t \frac{\sum_j P (T_j | Set) \; P (W_t | T_j)}{P (W_t)}$$

Model for P(Set)

- Model set probability to avoid inconsistent topics
  - Likely topic pairs, e.g.,
    - Clinton with Election Politics
    - Clinton with NAFTA
  - Unlikely topic pairs, e.g.,
    - Election Politics with NAFTA
- Approximate joint $P(T_1...T_M)$ as the product of all pairs normalized by the number of pairs to avoid bias

$$P(T_1...T_M) \approx \frac{\prod_{i=1}^{M-1} \prod_{j=i+1}^{M} P(T_i, T_j)}{\binom{M}{2}}$$

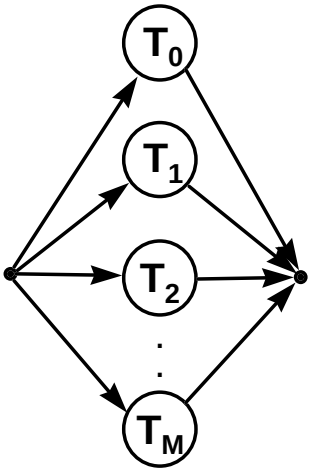
Estimation

- The count for a word is distributed among the labeled topics for a story in proportion to the posterior probability for each topic.

$$C(W_t | T_k) = \frac{P(T_k | Set)\ P(W_t | T_k)}{\sum_j P(T_j | Set)\ P(W_t | T_j)}$$

$$P(T_k | T_k \in Set) = \frac{\sum_i C(W_i | T_k)}{\text{\# words in stories with } T_k}$$

$$P(W_i | T_k) = \frac{C(W_i | T_k)}{\sum_i C(W_i | T_k)}$$



Search Algorithm (check it outXXX)

- Considering all possible topic sets is impractical.
- First, we consider each topic independently.
- Then, we score subsets of the most likely few topics.

Problem: When scoring one topic, words in the story from other topics will be very unlikely.

Solution: Only use positive information

$$\frac{P(W_t \mid T_j)}{P(W_t)} \approx f\left[\frac{P(W_t \mid T_j)}{P(W_t)}\right]$$

$$f[x] = \begin{cases} \theta, & x < \theta \\ x, & x \geq \theta \end{cases}$$

Topic Probabilities with New Method

- Given that a topic is in the set of topics for a story, we estimate the expected percentage of words that are related to that topic.
- Most topics are related to only 1-8% of the words
- 93.5% of the words are General English words

Topic	P (T | Set)
General English	0.935
Music, Black	0.085
...	
Politics and government	0.018
Clinton, Bill	0.020
Politics and government	0.018

Word Probabilities with New Method

- Observation probability of the relevant words is raised
- Probability of irrelevant words is greatly reduced

T = "Clinton, Bill"

Old Method

Rank	Word	P(W | T)
1	president	0.013
2	go	0.011
3	think	0.010
4	Clinton	0.009
5	say	0.008

New Method

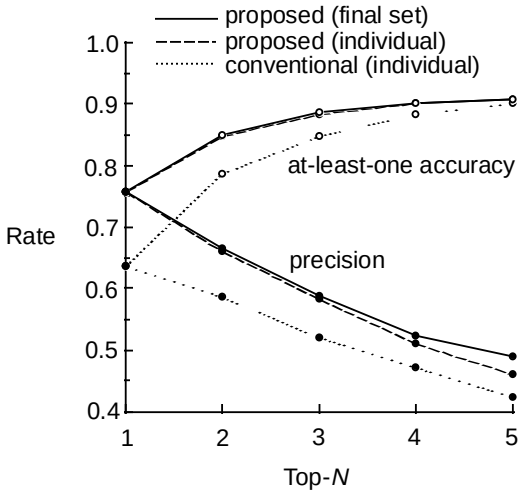
Rank	Word	P(W | T)
1	president	0.104
2	Clinton	0.096
3	house	0.036
4	white	0.034
.	.	.
.	.	.
36	go	0.003
44	think	0.003

Results

- Trained on 1 year of stories from July '95 to Jun '96 (42,502 stories)
- Tested on 989 stories from July '96
- Allowed 4,627 topics that occur at least twice
- OOT (out-of-topic) rate was 2.45%
- Results:
  - 75.8% of the first choice topics are among the annotated labels
  - 63.6% for the traditional method
- On cursory examination of errors, often the recognized topic was correct and the annotator failed to include it.

Recall / Precision Tradeoff

- **Recall: Fraction of annotated topics that are among the first N topics found**
- **Precision: Fraction of first N topics found that are among the annotated topics**



- **Recall and precision are always better for the new method.**
- **The co-occurence model of topics increases precision.**

Summary

- **Developed a new method of topic classification**
- **The model is more realistic**
  - **stories have several topics**
  - **only a few of the words are related to those topics**
- **The method is capable of topic classification among thousands of topics**